



An Empirical Study of Tagging for Personal Information Organization: Performance, Workload, Memory, and Consistency

Qin Gao

To cite this article: Qin Gao (2011) An Empirical Study of Tagging for Personal Information Organization: Performance, Workload, Memory, and Consistency, International Journal of Human-Computer Interaction, 27:9, 821-863, DOI: [10.1080/10447318.2011.555309](https://doi.org/10.1080/10447318.2011.555309)

To link to this article: <https://doi.org/10.1080/10447318.2011.555309>



Accepted author version posted online: 27 Jun 2011.
Published online: 25 Jul 2011.



Submit your article to this journal [↗](#)



Article views: 432



View related articles [↗](#)



Citing articles: 6 View citing articles [↗](#)

An Empirical Study of Tagging for Personal Information Organization: Performance, Workload, Memory, and Consistency

Qin Gao

Tsinghua University, Beijing, China

Tagging allows users to organize their information and retrieve it later with multiple, freely chosen keywords, which is impossible with categorical folders. The need to organize information for personal later retrieval has been found to be one of the most important motivations for tagging. Despite the popularity of the concept, more empirical evidence is still required to verify the real benefit of tagging for information organization and retrieval. Furthermore, the problems of inconsistency in tagging hamper the usefulness of tagging as an effective organization tool. The current study aims to investigate users' motivation, performance, and workload when they use tagging to organize personal information and how the system design could improve the process. First, a pilot study combining think-aloud and interviews was conducted to obtain insights on why and how users select tags. Then, the first experiment with 40 participants was conducted to empirically compare the performance and workload difference in information organization and retrieval tasks between categorization and tagging interfaces. The results show that tagging users reported a significantly higher level of mental demand and frustration when performing organizational tasks and a significantly higher level of temporal demand and error rate when performing retrieval tasks compared with categorization. However, tagging users tend to have better memory of the organized content. The second experiment aimed to study how individual tagging consistency can be improved by the proper visualization of tag suggestions. The impact of frequency visualization by font size and semantically clustering was studied with 40 participants. The results show that semantically clustered tag clouds improve tagging consistency significantly; when a semantic clustering effect is presented, frequency visualization by font size can significantly alleviate the physical demand perceived by users.

1. INTRODUCTION

Searching for information, organizing it, and finding it again later are important tasks in people's lives and work today. To retrieve and use useful information in

Correspondence should be addressed to Qin Gao, Department of Industrial Engineering, Tsinghua University, Beijing, 100084, P.R. China. E-mail: gaoqin@tsinghua.edu.cn

the future, the information must first be handled and organized properly by the individual. Previous studies found both the organization and the retrieval of personal information are not easy, and users' requirements to organize their personal files, e-mails, or web pages in a desired way have not been sufficiently satisfied by the primary solution available on most computers, the hierarchical categorization systems (Jones, Dumais, & Bruce, 2002; Ravasio, Schär, & Krueger, 2004). For information management, a major problem with such systems is the act of classification itself into a single category is often cognitively difficult (Lansdale, 1988; Malone, 1983) and maintaining a hierarchy requires continuous efforts from users (Nardi & Barreau, 1997; Ravasio et al., 2004). For information retrieval, with hierarchical folder systems it is difficult to link relevant information or to recall the context under which the information is organized (Gonçalves & Jorge, 2004; Ravasio et al., 2004). Thus users need to rely on their memory of the exclusive position of the piece of information in the hierarchy, which may not reflect the perspective they currently have for the information.

To address these drawbacks and problems, some other alternatives for information organization and retrieval have been proposed by system designers. Whereas some systems, like *Data Mountain*, continue to elaborate and augment the desktop metaphor (Agarawala & Balakrishnan, 2006; Mander, Salomon, & Wong, 1992; Robertson et al., 1998), some other systems abandon this metaphor and adopt a multiattribute-based approach, which utilizes useful document attributes generated or discovered by the system automatically to ease the retrieval of information (Adar, Karger, & Stein, 1999; Dourish et al., 2000; Fertig, Freeman, & Gelernter, 1996; Toda, Kadaoka, & Oku, 2007). Recently, tagging, a new approach relying on user-generated attributes, has emerged as a potential alternative for personal information management. Tagging was first developed to "harness the wisdom of crowds" in online content sharing applications like del.iciou.us and Flickr (<http://www.flickr.com>). Then the concept extended to proprietary information systems like enterprise information systems (Ahn, Brusilovsky, & Farzan, 2005; Braly & Froh, 2006; Grudin, 2006; Macgregor & McCulloch, 2006) and personal information management, like documents on one's hard drive (e.g., the tagging function in the newest operation systems like Microsoft Windows 7 and Mac OS X) and personal e-mails (e.g., the label function of Gmail).

A big advantage of tagging claimed by its proponents is that tag selections do not require exclusive decisions, and the cognitive cost of labeling objects might be reduced (Macgregor & McCulloch, 2006; Sinha, 2005) because multiple tags can be added to a single object, and there are no hierarchical relationships in tagging systems. Furthermore, the multiplicity of tag selection provides more "routes" to the desired information in later retrieval. According to such arguments, tagging could reduce the workload of information organization tasks and improve user performance on information retrieval tasks. However, there are also doubts about the real benefits of tagging for information organization and retrieval tasks. In particular, the flexibility of tagging usage, the fact that users are untrained in indexing, the lack of vocabulary control, and the ambiguous nature of language can lead to inconsistency in tagging (Golder & Huberman, 2006; Guy & Tonkin, 2006; Mathes, 2004), and this may affect the efficiency of later information retrieval. Some research found that tagging users were concerned with possible

confusion and delay in later information retrieval caused by inconsistent information organization behaviors (Civan, Jones, Klasnja, & Bruce, 2009; Quan, Bakshi, Huynh, & Karger, 2003), and they have a strong inclination to select tags they used before and to develop their own tagging rules (Rader & Wash, 2008; Wash & Rader, 2007). These studies, however, are mainly exploratory. To understand the real performance and workload difference between tagging and categorization, more empirical evidence is needed from experimental studies. Furthermore, designers are interested in how system design features could influence tagging consistency. Showing tag suggestions was found to be an effective way to shape people's tagging selection (Binkowski, 2006; Sen et al., 2006). As the number of tags increases and the load of visual search in tag lists or clouds becomes heavier, the presentation of tag suggestions may influence users' selection of tags and their consistency in tagging. However, this effect has not yet been examined.

The first goal of our study is to compare empirically the impacts of tagging versus categorization on individual users' performance, workload, and memory for information organization and retrieval tasks. The second goal is to investigate how the individual consistency of tagging can be improved by visualizing tag suggestions when the users select tags. Prior to the experimental studies, a pilot study combining interview and think-aloud was carried out among six participants to obtain insights of how Chinese users tag items and how they perceive the difference between tags and hierarchical categories. This was done because most previous studies of tagging motivation and behaviors investigated only Western users. Then two experimental studies were carried out. The first experiment compared tagging and categorization with regards to information organization and retrieval tasks and involved 40 participants. In the second experiment, the impact of frequency visualization by font size and semantic-similarity visualization by clustering on users' tag selection consistency and workload were tested, with another 40 participants. The findings are expected to enhance our understanding of tagging behavior and provide meaningful implications for system designers.

2. LITERATURE REVIEW

2.1. Information Organization With Hierarchical Categorization and Tagging Interfaces

Folder systems based on hierarchical categorization have been used for centuries to organize information, ranging from paper to electronic documents. Such systems follow a "put this there" model in the physical world, often resulting in a taxonomic categorization of documents. Previous research has explored the use of folder systems for paper-based documents, electronic files, e-mails, and web pages (Boardman & Sasse, 2004; Jones et al., 2002; Malone, 1983; Whittaker & Sidner, 1996). It has been repeatedly found, however, that users often have difficulties with hierarchical folders. Dumais and Jones (1985) identified two major problems with the use of hierarchical categorization: the difficulty of generating category names that will be used unambiguously and the fact that information in the real world falls into several overlapping and fuzzy categories. A hierarchical folder

system allows only one way to organize information, even though users may have several equally natural ways of organizing it. Above all, there is often a need to select a category based on what the document is or what the document is used for (Kwasnik, 1991). Correspondingly, Barsalou (1991) proposed that taxonomic and goal-oriented categories are suggested as two complementary ways to categorize the world. The latter refers to ad hoc categories that are collections of taxonomy kinds to integrate the information that people need to have to achieve the goal. Similarly, Ravasio et al. (2004) suggested that personal information can be viewed and organized from at least three different perspectives: content oriented, task oriented, and context oriented. Furthermore, even the subject of the information itself may be described through multiple facets (Janecek, 2007). Hierarchical folder systems, however, allow only one perspective and often lack support for goal-oriented or context-oriented views. This limitation was illustrated by the problem with folder hierarchies in the study of Jones, Phuwanartnurak, Gill, and Bruce (2005), in which a tension between organizing project information for current use and later reuse was found from participants. As a result, categorization implies a major cognitive effort and users tried to defer it as long as possible (Malone, 1983; Ravasio et al., 2004). Sometimes the job of information organization costs more time and effort than the information was worth (Barreau & Nardi, 1995; Lansdale, 1988). Even if an elaborate categorization scheme is developed, it still takes considerable effort to maintain the hierarchy to keep it up-to-date (Abrams, Baecker, & Chignell, 1998; Nardi & Barreau, 1997; Ravasio et al., 2004).

An important advantage of tagging is that it relaxes the limitation of one-to-many mapping with categorizations, allowing many-to-many mapping. Both taxonomic tags like "book," which describe what the document is, and ad hoc tags like "toread," which describe the way the document is used, can be added to a single document simultaneously (Veres, 2006). Lansdale (1988) suggested that using multiple keywords does not require the user to make a difficult decision as to which categorization to use, and this may reduce the cognitive effort needed for information organization. The low cognitive effort required was once considered an important contributor to the success of collaborative tagging systems (Macgregor & McCulloch, 2006; Wu, Zhang, & Yu, 2006) and is supported by several earlier qualitative studies. In the study of Civan et al. (2009), nine out of 10 participants stated that organizing with folders required more cognitive effort. This finding is consistent with a study on multiple categorization, in which more participants believed that maintaining an exclusive hierarchy required greater cognitive effort compared with multiple categorizations (Quan et al., 2003). Except for self-reported comments, however, neither study provided empirical validation for the low workload of tagging for information organization tasks.

The authors considered that although the statement that tagging needs less cognitive effort sounds intuitive, it may overlook the complexity of the cognitive process behind tagging. An underlying assumption of this statement is that users will put down every term coming to their mind as tags simply because the system allows them to do so. This assumption, however, may not be true, according to studies about the tagging motivation of individuals. The need to get one's own stuff organized and findable is a major motivation for tagging from the individual's point of view even in online content sharing communities (Ames &

Naaman, 2007; Marlow, Naaman, Boyd, & Davis, 2006; Sen et al., 2006). The users are found to be concerned with the ease of retrieval when they give the tag (Wash & Rader, 2007) and trying to maintain a certain consistency of tagging by reusing used tags or creating personal tagging rules or conventions for themselves (Civan et al., 2009; Rader & Wash, 2008; Sen et al., 2006; Wash & Rader, 2007). As the amount of information becomes intensive, the increased number of tags and rules may put great stress on users. In the study of Civan et al. (2009), five of 10 participants decided against many-to-many mapping when using labels, even though the system allows this, due to concerns of possible confusion, redundancy, and inefficiency. Furthermore, the visibility of part/whole or general/specific relationships in a folder hierarchy is considered important for information organization. In Jones et al. (2005), 13 of 14 participants would not give up using folders even if they were given the option of finding information with Google like search, and one important reason is the visibility of the relationships among items that are embodied by folders. Civan et al. (2009) found that some participants used workarounds, such as giving a common prefix to multiple folders/labels in order to incorporate whole-part and general-specific relationships into their organization scheme, when they found that the testing systems did not explicitly support hierarchy.

2.2. Information Retrieval With Hierarchical Categorization and Tagging Interfaces

A big difference between categorization and tagging is that the former provides a sense of digital location of an information item, whereas the latter does not. The advantage of spatial cues for the ease of information retrieval, however, should not be taken for granted. Although some earlier psychological research seems to suggest that locational attributes of information can be remembered incidentally and effortlessly (Hasher & Zacks, 1979; Mandler, Seegmiller, & Day, 1977), later studies that involve more practical information organization tasks found little evidence to support the effectiveness of solely spatial cues for information retrieval. For example, the Jones and Dumais experiments (Jones & Dumais, 1986) show that the usefulness of location-based approach to information retrieval is very limited, and enrichment of the spatial cues do not produce any appreciable benefits. Lansdale's (1991) research further found that when the location information of documents is only arbitrary associations to other items, users' recall for location is poor; when the location information is ascribed meaning during the filing process, there is a clear advantage in recall of location information. These results suggest that the location of a document in one's folder hierarchy says much more than itself; it may express the nature of that document, how it is evaluated against criterion for creating the hierarchy, and how it fits into the preexisting system. In Civan et al.'s (2009) study, users appreciated this feature of the folder system, for it provides more visual cues and allows them to use spatial memory for information refinding. Barreau and Nardi's research (Barreau & Nardi, 1995; Nardi, Anderson, & Erickson, 1995) also found that computer users prefer location-based search over direct search by keywords to refind files, partially due to the difficulty

to remember the name of files. Ravasio et al. (2004) found similar preferences but attributed it to the poor usability and usefulness of the built-in search tools.

Retrieving information using the way-finding approach, however, is influenced by users' spatial capability and working memory capacity (Pak, Rogers, & Fisk, 2006; Stanney & Salvendy, 1995). Especially when the depth of the hierarchy of information architecture increases, the navigation problem becomes more and more treacherous, resulting in longer performance times, higher error rates, and higher perceived complexity (Jacko & Salvendy, 1996; Kiger, 1984; Miller, 1981; Seagull & Walker, 1992; Snowberry, Pakinson, & Sisson, 1983; Zaphiris, 2000).

Another difference between categorization and tagging is exclusivity versus multiplicity. The exclusivity of a categorization scheme assures that a piece of information is in one stable place. This supports systematic and exhaustive search (Civan et al., 2009). In the study of Quan et al. (2003), participants reported that organizing information with taxonomic categorization requires less scanning to locate relevant categories compared with multiple categorization. In contrast, tagging users may benefit from the multiplicity of tagging by means of multiple retrieval routes (Civan et al., 2009). The multiplicity of tagging also enables interlinking of related information. Gonçalves and Jorge (2004) noted that when describing a particular document, users often made small descriptions of other related documents. Ravasio et al. (2004) found that establishing linkage between related information is found an expressed need for electronic information organization. Through common tags shared by different objects, tagging allows each object linked with others in many different ways.

2.3. Influence of Tagging Versus Categorization on Memory

Information organization inherently involves mental operations such as perception, comprehension and elaboration, and these operations are believed to affect users' memory of the information (Anderson, 2000; Craik & Lockhart, 1972). The organization of information in memory, the level of processing, the degree of elaboration of the material, and the match between organization and retrieval tasks were found to influence people's memory (Anderson & Reder, 1979; Bradshaw & Anderson, 1982; Bransford, Franks, Morris, & Stein, 1979; Craik & Lockhart, 1972; Freeman, Romney, & Freeman, 1987). Categorization and tagging each have their own advantages and disadvantages regarding memory development. On one hand, the hierarchical structure of categorization may encourage or force users to form a clear mental structure of the information. Memory research has found that the degree of development of a mental structure is a principal factor in determining memory performance (Bower, Clark, Lesgold, & Winzenz, 1969; Mandler, 1967; Tulving, 1962). On the other hand, tagging users were often found using terms related to themselves (e.g., subjective tags, self-reference, task organization) in addition to categorical terms when tagging a piece of information (Golder & Huberman, 2006; Veres, 2006; Xu, Fu, Mao, & Su, 2006). This suggests that tagging users may process the information item more elaborately by generating more relations between the information and their own situations. Furthermore, there are

likely to be more similarities between the retrieval task and the organization procedure in tagging because taggers are allowed to incorporate multiple possible retrieval cues during the organization phase (Wash & Rader, 2007; Xu et al., 2006). In the study of Budiu, Pirolli, and Hong (2009), participants recalled more facts in tagging conditions than in the no-tagging condition in later sessions. However no study has compared users' memory in tagging and categorization conditions.

2.4. Tagging and Indexing Inconsistency

A major problem with tagging is the inconsistency in tag selections between and within individuals (Civan et al., 2009; Golder & Huberman, 2006; Wash & Rader, 2007), due to the absence of a controlled vocabulary and the ability of multiple categorizations. In addition to "bad" tags like misspelled tags, badly encoded tags, tags that do not follow conventions in issues such as case and number, and mixed-use of plurals and singulars (Guy & Tonkin, 2006), Golder and Huberman (2006) identified three types of semantic inconsistency that is inevitable in free tagging systems: polysemy (one term related to many meanings), synonymy (multiple terms related to the same meanings), and basic-level variation (terms at different specificity levels used arbitrarily for one object). From the system perspective, such inconsistency may have some merits. Some researchers have argued that it is precisely the lexical ambiguity of tagging that allows a true representation of knowledge and multiple interpretations of the same content and enables users to benefit from other people's discoveries (Campbell, 2006; Golder & Huber, 2006; Shirky, 2005). Some other research found that consistent tagging patterns emerge as the number of tags and documents increases and that inconsistencies may follow several predictable patterns (Golder & Huberman, 2006; Kipp & Campbell, 2006).

For individual users who tag information for later retrieval, however, tagging consistency may be more desirable. For the purpose of information organization and retrieval, tagging is similar to indexing. Indexing research has found that indexing consistency is a major dimension of information quality and is desired for effective later retrieval (Borlund & Ingwersen, 1997; Hurwitz, 1969; Naumann & Rolker, 2000; Soergel, 1994). The results of Leonard's experiments showed a trend toward a moderate to high association between indexing consistency and retrieval performance (Leonard, 1975).

Recent tagging research revealed two characteristics of tagging behaviors in relation to tagging consistency. First, taggers users were found to have a strong intention to reuse tags they used before (Rader & Wash, 2008; Sen et al., 2006; Wash & Rader, 2007). Through interviews of 12 experienced tagging users, Wash and Rader (2007) found reusing tags the user had applied before and adhering to mental rules or definitions for specific tags were two main heuristics for selecting tags. The regression result of Rader and Wash (2008) on del.icio.us data confirmed the significant influence of personal used tags on users' tag selection, and its impact is larger than the impact of tags used by other people. Several other studies reported that tagging users tend to create personal rules or definitions for specific tags to regulate their tagging behaviors (Golder & Huberman, 2006; Kipp & Campbell,

2006; Wash & Rader, 2007). However, these studies also show that users often fail to apply these rules consistently as the number of tags and rules increases. In the study of Civan et al. (2009) on the use of tagging in e-mail organization tasks, half of the participants even gave up the multiple categorization capability of tagging due to concerns of possible confusion and redundancy caused by multiplicity.

2.5. Tag Suggestion and Visualization

Some research has investigated the effect of system feature design on how users give tags and found that providing tag suggestions has a significant impact on users' choice of tags (Binkowski, 2006; Sen et al., 2006). The results of Binkowski's experiment shows that users' selections of tags will be more similar to each other if they are provided a list of the 10 most popular tags added to the website (Binkowski, 2006). In another study (Sen et al., 2006), it was found that showing a list of preused tags, assigned by oneself or by other taggers, affects future tag selections for movies.

As the amount of information and the number of tags increases, it will become more and more difficult for users to browse all used tags to select the proper ones. How the system selects and displays suggested tags will thus tagging selection. Proper visualization techniques will make the browsing process easier and relax the demands on the performance of retrieval and query generation. With proper visual aids, users will be able to apply perceptual inference, in addition to logical inference, to discover information patterns or relationships and to recognize relevant documents (Lin, 1997).

Tag clouds are widely used to display a set of tags in tagging sites. In a tag cloud, attributes of the text, such as size, weight, or color, are used to represent features, such as frequency, of the associated terms (Rivadeneira, Gruen, Muller, & Millen, 2007). Whereas the first generation of tag clouds are frequency visualization, the second generation of tag clouds focuses on revealing semantic relations between the tags (Nielsen, 2007). It is assumed that visualizing high-level semantic relationships among tags, rather than low-level variables like frequency of use or alphabetic ordering, will make tag clouds a more useful tool for users (Choy & Lui, 2006). Hassan-Montero and Herrero-Solana (2006) proposed an approach to build semantic clusters of tags based on their co-occurrence similarity. Likewise, Choy and Lui (2006) proposed to use Latent Semantic Analysis to evaluate the tag similarity, and to use the result to train a self-organization map. There are also research efforts to extract hierarchical structure from the flat tag space and visualize it in a tag cloud (Brooks & Montanez, 2006; Li, Bao, Yu, Fei, & Su, 2007). Schrammel et al. (2009) examined the effect of semantically clustered tag clouds on users' performance in search tasks, in comparisons with alphabetically ordered tag clouds and random tag clouds. It was found that semantically clustered tag clouds provide improvements over random layouts in specific search tasks and that they tend to increase the attention paid to tags in small fonts compared to other layouts. They are also preferred in general search tasks. However, the impact of semantically clustered tag clouds on users' selections of tags has not yet been studied.

3. PILOT STUDY

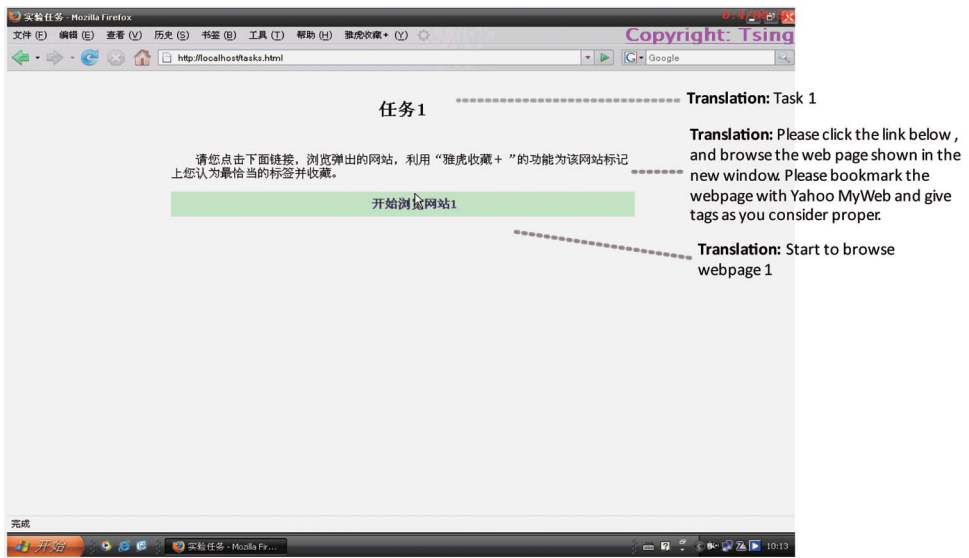
3.1. Methods

To obtain insights into users' intentions, perceptions, and behaviors in the tagging process, a pilot study combining think-aloud tests and semistructured interviews was conducted. The goal of the pilot study was to understand how individual users—in particular, Chinese users—use tagging for information organization and retrieval tasks, and why and how they tag.

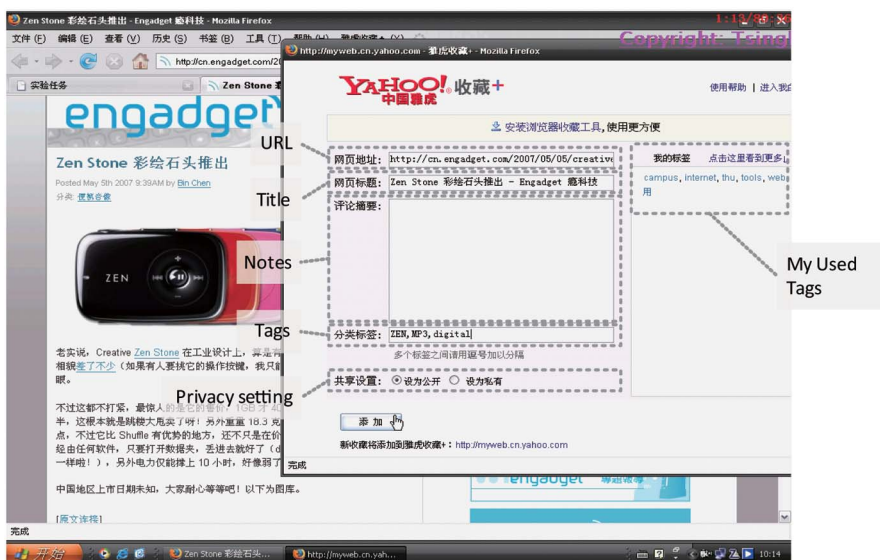
We conducted six 1- to 2-hr semistructured interviews with Chinese tagging users. The recruitment was carried out by posting on the public Bulletin Board System (BBS) of Tsinghua University. Six participants were recruited, aged 21 to 29 years, all male, including three undergraduate students, two master's students, and one information technology professional. Their average computer experience was 8.5 years, and their average online time was 6.67 hr per day. All had used tagging for more than 3 months. Four were experienced tagging users with more than 2 years of tagging experience, posting more than 10 bookmarks in social bookmarking systems per week. The other two were novice users who had been using tagging for less than 1 year and only used tagging in blogs and photo-sharing websites. The sample represents a young, highly educated, and tech-savvy population, and both novice users and experienced users were included to represent a wide variety of usage patterns.

Each interview commenced with the instructor introducing the purpose of the study and demonstrating the use of the selected tagging application—the Yahoo! MyWeb¹ Firefox browser plug-in. Yahoo! MyWeb was selected because it has similar functions to del.icio.us but is in Chinese, and the plug-in enables users to tag a web page without leaving the current position. Then participants were sequentially shown 15 web pages in Firefox browser on a Lenovo Thinkpad R51 notebook (Lenovo, Beijing, China) and gave tags to each web page. Figure 1 shows the task instruction screen and the tagging task screen. The web pages were selected from the recently bookmarked websites in Yahoo! MyWeb, so that the content should be interesting to the general public and they should have sufficient understanding of the content to allow meaningful tagging. The topics included news, digital devices, travel, and so on. This manipulation may differ from the real use pattern, in which users select materials to tag on their own. However, using controlled materials allowed us to compare the different perspectives of different users on the same web page. All the web pages were retrieved from the Internet and stored on a local server. A new MyWeb account was created for the study, and the researcher cleaned the history of the account before each session. The participant started tagging with an empty account. The sequence of 15 web pages was randomized. Participants were encouraged to vocalize their thoughts, opinions, and feelings while tagging these pages. Upon completion of all tasks, the participants were interviewed about their motivation to use tagging tools, the perceived difference between tagging can categorization, the typical situations they use tagging for

¹Yahoo! MyWeb was shut down in February 2009, after Yahoo!'s acquirement of del.icio.us.



(a) Task instruction screen



(b) Tagging screen

FIGURE 1 Screenshots of the pilot study (color figure available online).

information organization or retrieval, how they select tags, and how they manage tags. Finally, the participants were given a list of titles of web pages they had tagged in the first phase and asked to recall the tags they had added in the first phase. Each session was video- and audio- recorded. The records were transcribed and coded.

3.2. Findings From the Pilot Study

Tagging motivation. All participants reported that they tagged mainly for personal benefit, and four of them reported that there were times they wanted to share with others and they would tag differently when considering sharing with others. They would choose more general words when tagging for others, avoiding using ad hoc definitions in tagging. Only one participant (P2) reported he tagged for others on a regular basis. He also mentioned that he would try to guess the possible keywords other people would use if they want to find the content. This participant used the most tags among all participants, tagging each web page with four to seven tags.

From the personal perspective, participants reported two major functions of tagging: facilitating future retrieval and annotating content. Most participants were aware that they had used tags for both functions: "Some tags are for retrieving, and some tags are just for reading (to get an idea what the item is about)." Two participants explicitly indicated that they use tags as both categories and keywords, but they also admitted that such awareness comes from post hoc reflections and that they did not deliberately distinguish these two types of tags in daily use. Participants exhibited different tagging patterns and concerns when they selected tags for different functions.

Tagging for future retrieval. Four participants felt that using tags to refine information in one's own collection can improve the accuracy of retrieval compared with categorization and general keyword search. However, it depends on how well the user can remember or guess the tags he or she used. Two approaches to facilitate later retrieval were found in participants' tagging behavior: increasing tag consistency and increasing tag redundancy.

All participants agreed that they would reuse old tags as much as possible in order to improve the tag consistency. Participants also created ad hoc definitions or personal rules. Except for one novice user, five of the participants had some tags with personal definitions. For example, one user clearly distinguished between "media" (referring to print media and journalism) and "audio" or "video" (referring to computerized rich media content); another user used "fun" to refer to interesting but nontechnological content (e.g., jokes) and used "interesting" to refer to interesting technological content; two users deliberately limited the level of specificity (in both cases, avoiding using too specific tags) in their tags. Three participants emphasized the importance of such definitions despite their imperfections: "I know this kind of definition is strange for others. But it is the way that works for me." Such rules may be often incomplete and

ill-defined. For example, terms of different categorical levels were mixed together (e.g., a participant defined “politics,” which actually referred to “domestic politics” in his definition system; he used it at the same level with “international politics”). Rules sometimes conflict with each other, and few participants can apply them consistently. In all think-aloud sessions in which personal rules were mentioned, there were occasions when the participant found that his later tag selection was conflicting with some personal rules he had mentioned before. For example, one participant (P1) reported that he would use high-level categorizing words as his own as tags in general in order to reduce the number of different tags he has. When he was given a web page about a product, a pocket DJ with a mixer, he was found clicking and copying descriptive words from the web page and using them as tags. He realized the inconsistency then and admitted that his tagging pattern is more arbitrary than he had previously believed.

Although all participants were aware of the inconsistencies in their tags, most participants were reluctant to “clean” their tags once they had created them. Only one participant (P5), a novice user, modified his tags immediately after being reminded of the conflicts. All the experienced tagging users considered it impossible to apply all their rules consistently and simply accepted this as a fact. One participant reported that he would fix some major inconsistencies in his tag sets if he found it necessary. The rest said that they never or very seldom make changes to used tags, but two of them said they would like to do such work if the tagging system provided convenient tools to do so.

Increasing tag redundancy was another approach to facilitate future retrieval, often as a remedy to the failure of consistent tagging. Similar tags were used simultaneously to increase the probability of being found later. For example, a participant reported that 90% of the objects tagged with “product” and the objects tagged with “design” in his collection overlapped, yet he often added both tags to a single content to ensure that he can retrieve the content easily. Two participants found that redundant tags are effective when their memory of the object and of the tags is vague.

Tagging for annotating the content. Participants found tagging as a way to express their understanding or summary of the content: “Special tags like ChildMp3 and ToyMp3 described the characteristics of the web page content. Such tags help me to have an idea of the content in the list view very quickly, without clicking through each link.” Another participant called such tags “tags for reading.” Two participants found that such tags were also useful to indicate the value/interestingness of the tagged content and attach personal perspectives to the content when he wanted to share it with others. Most participants considered that they were not likely to use such tags for retrieval because they may forget the exact tags.

Criteria for selecting tags. By analyzing both the think-aloud test and interview results, we found the following criteria for selecting “proper” tags:

- **Likelihood of being used for organizing more information later:** Tags that are likely to be added to other information later on are preferred. One participant said, "When I give tags, I hope they can be used later on. If I find a tag only added to one content long after its creation, I will think it is useless and try to delete or combine such tags with more popular ones. But I don't have time to do such [cleaning] work regularly." Another participant considered the frequency of usage of tags as showing his current interests, and he would take a look at his overall use of tags from time to time.
- **Likelihood of being used for retrieving current information later:** All participants reported that they thought about how the information will be used (e.g., for temporal project, long-term interest, general reference) when they selected tags. Concerns about the context in which the information will be retrieved had a direct impact on their tag selections. A participant reported that he thought about the most possible words he would use to retrieve the information later when giving tags.
- **Centrality and descriptiveness of the content:** Most participants felt that tags that are central and descriptive to the content necessary, even though the tag may not fit their organizational schemes: "I didn't have [Athens (location)] before, and I don't think I will add a lot of content to this tag later. But it is too important for describing the content properly."
- **Minimizing effort to generate and remember tags:** Several participants were found to use different measures to minimize the cognitive effort of generating and remembering tags. Two participants copied and pasted words from the web page as tags in some tasks, and three participants used keywords given by the author as tagging references. In particular, one of the three said that he searches for such keywords proactively before he gives tags. Participants were also found to minimize the effort of remembering tags by selecting words they are familiar with. When the topic of the web page was outside of their familiar vocabulary, they would try to use familiar words even when this led to imprecise descriptions. The concern of unfamiliar words was that they may not be able to tell what was related to that term later: "I must be able to understand at first sight what my tags mean and what is included in this tag."
- **Consistency with one's own tagging pattern:** As discussed previously, all participants had concerns about tagging consistency, and they used certain personal heuristics (e.g., reuse of tags, ad hoc tagging rules, limiting the level of specificity) to improve their tagging consistency.

There were often conflicts between these criteria. For example, words that were central to the content may likely not be used to tag other content that the user is going to collect. In the tagging process, they were found to use different criteria at different times. Although users want to maintain consistency in their tagging pattern, they were aware of the incompleteness/imperfections of their own tagging scheme: "Though I have some tagging traditions for myself, my tagging process is still arbitrary in many cases. Sometimes keywords just jump into my brain after I read the content, and I may use some of them as tags. Even I myself cannot explain why I select a word as a tag but not another sometimes." Yet they were reluctant to

refine their rules. Two of them reported that they try to use a few “big” tags so as to improve the consistency of tagging, but they also found that this results in long lists under each “big” tag, which reduces the efficiency of information retrieval.

4. EXPERIMENT 1

4.1. Research Questions

Qualitative results from the pilot study found that tagging users have a number of criteria for selecting tags. This may develop more memory traces of the content, but it also leads to conflicts among different criteria. In information retrieval tasks, the multiple possible paths back to an item allowed in tagging may facilitate the retrieval. However, flexible and inconsistent tagging behaviors may impede effective retrieval. The results suggest that there may be differences in users’ performance, workload, and memory between tagging and categorization interfaces, but more valid evidence is needed to verify it. The goal of Experiment 1 was to empirically compare the performance, workload, and memory differences between tagging and categorization interfaces. The three research questions of Experiment 1 were as follows:

- What is the impact of tagging compared to categorization on user workload during information organization tasks?
- What is the impact of tagging compared to categorization on task performance and user workload during information retrieval tasks?
- What is the impact of tagging compared to categorization on users’ memory of the content being organized?

Differences in retrieval performance were of greater interest than organization performance for several reasons. First, different people have their own individual organization preferences, and their styles may work just for themselves. Thus the retrieval performance reflects the quality of the organization scheme. Second, the completion time of organizational tasks used in the study of Pak (2007) may not be a proper measure. We instructed our participants to organize information at their own pace as they usually do. Participants may find certain content interesting and spend more time reading the content than was necessary to organize it, as found by Quan et al. (2003).

4.2. Methodology

Participants. Forty participants (17 female, 23 male) were invited to take part in the experiment. Participants were solicited by posting on campus BBS. Participants comprised undergraduate and graduate students at Tsinghua University. The average age was 22.6 ($SD = 1.30$). Participants had used computers for an average of 8.2 years ($SD = 2.77$) and had used the Internet for an average

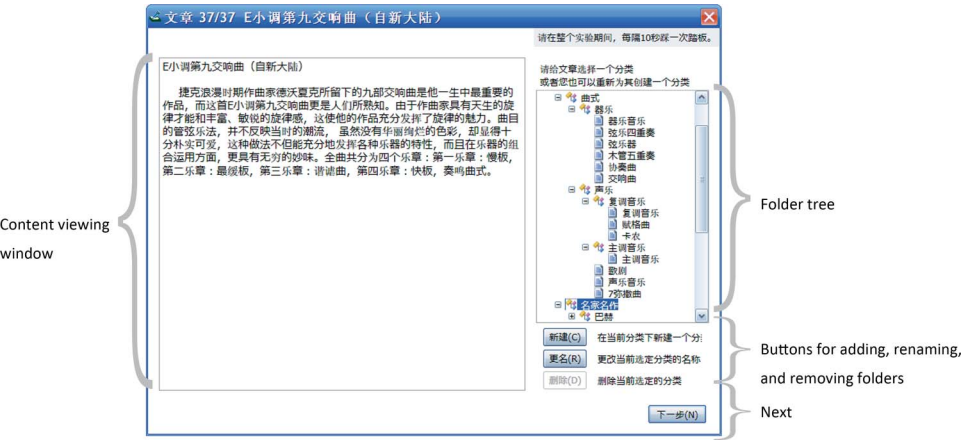
of 7.1 years ($SD = 1.55$). Overall, the participants constituted a representative sample of college students, with high levels of education and rich Internet experiences. They were randomly assigned into one of two experimental groups: the tagging group and the categorization group. An examination of age, computer experience, and Internet experience showed no significant difference between the two experimental groups.

Materials. Thirty-eight articles about Western classical music were composed based on information from Wikipedia and two music textbooks. They covered a set of diverse yet related topics, including historical musicology, musical styles, musical instruments, famous musicians, performances, and master works. Topics about Western classical music were chosen because most of our participants had a positive attitude toward Western classical music. None of them, however, had in-depth knowledge in this domain. Thus, the influence of previous knowledge was avoided. All articles were no longer than 500 Chinese characters (about 250 English words after translation) and were written in easy Chinese so as to be understandable for people with little previous knowledge.

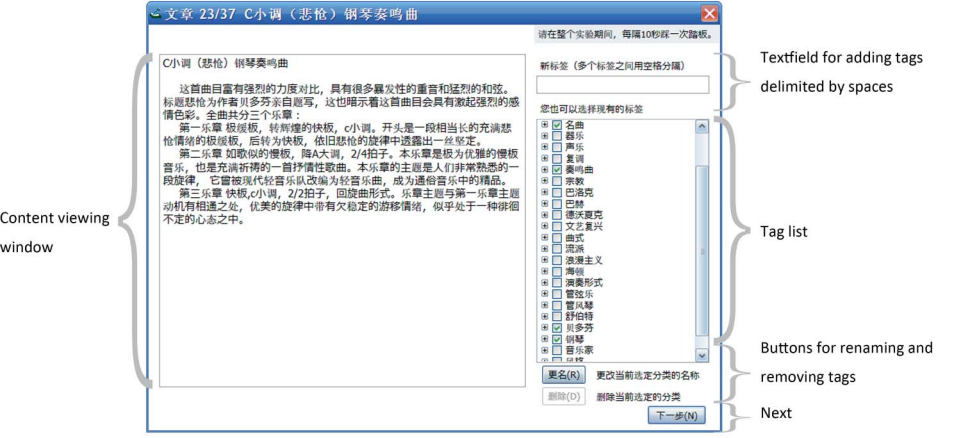
Testing systems. A testing system was created to with Microsoft Visual C++ and presented on a Lenovo T61P laptop, equipped with 15.4-in. LCD display set to 1280×800 pixels resolution. Figure 2 shows the categorization and tagging interfaces for organizing information. On the left is a content viewing window, showing the articles to be organized as well as the instructions. On the right is an organization panel using either categorical folders or tags. In the categorization condition, the panel displays all categorical folders in a hierarchical structure (Figure 2a). The titles of articles assigned to a category are also presented under the category in the tree structure but denoted by a different document icon. Child folders and articles can be hidden by either clicking the “+/-” icon to the left of each folder or double-clicking the folder name. The user can add, remove, or rename folders using the buttons at the bottom of the organization panel. The tag panel displays the tags (Figure 2b) in a linear list, each with a checkbox to the left. Users can assign multiple tags to an article by checking multiple boxes. Similarly, the titles of articles assigned to each tag are presented under the tag, and can be hidden by clicking the “+/-” icon or double-clicking the tag name. Adding new tags can be accomplished by typing tags, delimited by space, in the text field on top of the tag list.

In the retrieval session, the user had to browse his or her own collection built in the organization session in order to search for answers for the task at hand. Figure 3 shows the interfaces for retrieval tasks. The user could navigate the collection using the folder tree or tag list. All child folders and article titles were hidden at the beginning of each retrieval task. Double-clicking an article title would display the content of an article in the content viewing window.

Dependent variables. The five dependent variables were the completion time of the retrieval tasks, errors in retrieval tasks, task workload, memory of



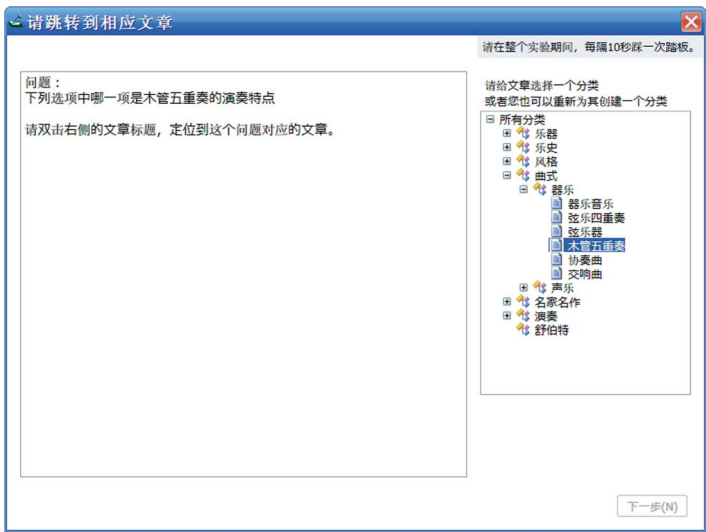
(a) Organizing articles with categorical folders



(b) Organizing articles with tags

FIGURE 2 Information organization with categorization and tagging interfaces in Experiment 1 (color figure available online).

the content, and user satisfaction. The completion time of each retrieval task was recorded by the testing system and the sum of all tasks was calculated as the overall measure. Errors in retrieval tasks were defined as the average number of excessive clicks required to complete a retrieval task. The number of excessive clicks was calculated by the difference between the number of clicks a participant performed to complete a retrieval task and the number of necessary clicks to complete the task. Memory of the content was measured by the score participants received in a postexperiment test. In the test, participants had to answer 19 multiple-choice questions based on the content used during the experiment, and the score was defined as the percentage of questions a participant correctly answered out of all questions the participant answered. Satisfaction was measured with a score obtained through a general satisfaction questionnaire of 14 items on



(a) Retrieving articles with categorical folders



(b) Retrieving articles with tags

FIGURE 3 Information retrieval with categorization and tagging interfaces in Experiment 1 (color figure available online).

a scale of 1 (*lowest satisfaction*) to 7 (*highest satisfaction*). The questionnaire was adopted from Cook (1991).

We used two methods to measure task workload: NASA-TLX and secondary task (time estimation) performance. NASA-TLX scale is a widely used subjective assessment and was selected for its good interrater reliability (Vidulich & Tsang,

1985), sensitivity (Liu & Wickens, 1988; Vidulich & Tsang, 1985), ease of rating, and unobtrusiveness. It comprises six subscales, measuring mental demand, physical demand, temporal demand, performance, effort, and frustration, respectively. Each of the six subscales was found to contribute independent information about the structure of different tasks (Hart & Staveland, 1988). The time estimation technique was selected due to its high sensitivity (Casali & Wierwille, 1984; Liu & Wickens, 1994). During the experimental tasks, participants were instructed to estimate constant time intervals (10 s) by stepping on a pedal connected to the computer while they carried out the primary task. The underlying logic is that spare capacity, not being directed to the performance of the primary task, will be used by the secondary task. Thus the performance of the secondary task can reflect the level of workload in the primary task.

Procedure. All experiments were carried out in the Usability Lab at Tsinghua University, and each participant was tested individually. The computer program automatically traced and timed (to the nearest one thousandth of a second) the participants' movements. Each experiment began with the instructor introducing the purpose of the study and demonstrating the use of the testing system. Then a 1-min calibration block was given in which participants watched a square on screen changing its color every 10 s. Following that, participants performed a 3-min baseline time estimation block with no concurrent activities. After completing all experimental tasks, participants performed another 3-min baseline time estimation block without concurrent activities. The average time estimation performance of the two blocks was used as the baseline time estimation of the individuals.

Before starting the organization session, a brief practice session with 10 trials was conducted to familiarize participant with both the operation of the testing system and the method of performing the two tasks simultaneously. In the following organization session, participants organized 38 articles with categorization or tagging interfaces. Articles were presented one at a time in a random sequence. The participants were asked to create an organization scheme from scratch. They were instructed to take as much time as they needed to organize information in a way they considered good enough and were allowed to go back and change categories or tags assigned previously. Then they were given the NASA-TLX questionnaire and satisfaction questionnaire for the organization session. Next, they were briefly interviewed about their information management habits, and demographic information was collected. We assume that this portion of the study could be considered as a distraction period and the recency effect could be reduced.

In the subsequent information retrieval session, participants were asked to find answers for 16 questions from their collections. Questions were presented one at a time to the participant in a randomized sequence, and each question asked about an issue discussed only in one article in the collection. To answer a question, the user had to navigate his collection and find the corresponding article. If the selected article was wrong, the user saw a pop-up window indicating the error and asking him to continue search until the correct article was found. The participant

was instructed to complete the retrieval task as fast as possible without sacrificing accuracy. Upon the completion of all tasks, NASA-TLX questionnaire and satisfaction questionnaire for the retrieval session were given. Following the second baseline time estimation block, the participant was asked to take an uninformed memory test. Finally, the participants were debriefed and given 50 RMB Yuan for their participation.

4.3. Results

Workload and satisfaction for organization tasks. Previous studies on NASA-TLX (Hart & Staveland, 1988; Liu & Wickens, 1994) show that the detailed analysis of the NASA subscales could provide insightful information about subjective workload, which may not be obtained from the weighted ratings. Therefore, both the means of the overall NASA-TLX score and the means of the six subscales were examined in this study. The overall workload score was calculated as a weighted combination of all subscales, where the weightings of the subscales are calculated from the pair-comparisons of the contributions of the subscales (Hart & Staveland, 1988). The rating of the importance of the six subscales is listed in Table 1. Performance, effort, and mental demand are the three most important factors contributing to the overall workload in both of the experimental sessions.

The difference in workload between the categorization and tagging conditions were tested by analysis of variance (ANOVA). As shown in Table 2, there is a marginally significant difference, $F(1, 39) = 3.27, p = .08$, in the overall workload: tagging users report higher overall workload ($M = 59.14, SD = 7.04$) than categorization users ($M = 54.54, SD = 8.95$) in information organization tasks. Further comparisons of the means of the six subscale ratings revealed significant differences in mental demand, $F(1, 39) = 4.73, p = .04$, and in frustration, $F(1, 39) = 4.66, p = .04$. The Mental Demand subscale measures how much mental and perceptual activity was required, and the Frustration subscale measures the extent to which the participant felt insecure, discouraged, irritated, stressed, and annoyed during the task. The results showed that tagging users experienced more demand of mental activities ($M = 67.38, SD = 12.15$) and more frustration ($M = 48.78, SD = 17.46$) than categorization users did (Mental Demand: $M = 58.05, SD = 14.84$; Frustration: $M = 35.55, SD = 21.12$). No significant difference was found in the other subscales. In addition, no significant difference was found in the satisfaction ratings.

Table 1: Sources-of-Workload Weight of NASA-TLX Subscales in Experiment 1

	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration
Organization session						
Sources-of-workload weight	3.16	1.184	1.868	3.710526	3.18	1.89
Rank order	3	6	5	1	2	4
Retrieval session						
Sources-of-workload weight	2.68	1.53	2.42	3.66	3.03	1.73
Rank order	3	6	4	1	2	5

Table 2: Comparisons of NASA-TLX Scores and User Satisfaction for Information Organization Tasks

	<i>Tagging^a</i>		<i>Categorization^a</i>		<i>F</i>	<i>p</i>	<i>Difference in %</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Workload							
Mental demand	67.38	12.15	58.05	14.84	4.73	.04	16%
Physical demand	43.50	26.12	48.38	28.55	0.32	.58	-10%
Temporal demand	44.50	16.67	44.63	16.77	0.00	.98	0%
Performance	45.28	12.28	41.00	18.50	0.74	.40	10%
Effort	71.15	11.41	72.00	11.17	0.06	.81	-1%
Frustration level	48.78	17.46	35.55	21.12	4.66	.04	37%
Global	59.14	7.04	54.54	8.95	3.27	.08	8%
Satisfaction	4.82	0.51	5.00	0.39	1.62	.21	-4%

^a*N* = 20.

Performance, workload, and satisfaction for information retrieval tasks and memory for the content. The completion time and the error rate of retrieval tasks were compared with ANOVA, as shown in Table 3. For the data that violated the normal distribution assumption of ANOVA, a nonparametric testing method (Wilcoxon two-sample test) for independent samples was adopted. No significant difference was found in the completion time, but there was a significant difference in the error rate (Wilcoxon two-sample test: $Z = -2.03$, $p = .02$). The error rate of tagging users ($M = 0.50$, $SD = .58$) is significantly higher than that of folder users ($M = 0.13$, $SD = .13$). Ideally, tagging users can retrieve every article with two clicks (one for the tag, one for the article), which is often lower than the number

Table 3: Comparisons of Information Retrieval Performance, Workload, Satisfaction, and Memory

	<i>Tagging^a</i>		<i>Categorization^a</i>		<i>F</i>	<i>p</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Completion time (seconds)	382.1	101.9	430.8	272.3	0.566	.46
Error rate	0.50	0.58	0.19	0.13	$Z = -2.03^*$.02
Total no. of clicks to complete tasks	44.60	10.97	43.85	6.45	$z = 0.69^*$.24
Workload	47.39	14.00	42.66	13.81	1.16	.29
Mental	52.58	20.77	48.45	19.07	0.43	.52
Physical	36.65	22.71	32.55	20.73	0.36	.55
Temporal	48.75	23.06	32.63	20.70	5.41	.03
Performance	32.50	18.69	30.00	22.58	$Z = -0.745^*$.48
Effort	57.88	16.65	52.88	17.42	0.86	.36
Frustration	33.25	17.05	29.05	22.10	0.45	.51
Memory score (%)	60.11	15.12	54.02	17.00	$Z = 1.40^*$.16
Satisfaction	5.15	0.44	5.13	0.32	.033	.85

^a*N* = 20.

*Wilcoxon two-sample test was used since the assumption of normality for analysis of variance was violated.

of necessary clicks to retrieve an article using hierarchical folders. However, our result showed that the total number of clicks a tagging user used for retrieval tasks ($M = 44.6$, $SD = 10.97$) was similar to that of a folder user ($M = 43.85$, $SD = 6.45$; $Z = 0.69$, $p = .24$). This is connected to the high error rate of tagging. Although tagging allows users to retrieve an article by short paths due to the flat structure of the tag space, our results found that there is no significant reduction in the number of clicks.

The examination of the NASA-TLX rating results showed no significant difference in the overall workload. A detailed analysis revealed no significant differences in the subscales except for the Temporal Demand ($F = 5.41$, $p = .03$). This subscale measures how much time pressure the user feels due to the rate at which the task occurred. Tagging users perceived more time pressure ($M = 48.75$, $SD = 23.06$) than folder users did ($M = 32.63$, $SD = 20.70$).

The comparison of memory test scores found that tagging users tended to have a better memory of the materials ($M = 11.50$, $SD = 2.82$) than folder users ($M = 10.20$, $SD = 3.16$; $Z = 1.40$, $p = .16$), although the difference was not significant. No significant difference was found in user satisfaction between groups.

Time estimation. For the secondary time estimation task, the median and the average absolute deviation of scores from the median have been found to be more representative measures of central tendency and dispersion than the mean and the standard deviation (Hart, 1975; Liu & Wickens, 1994). Thus they were chosen to describe the time estimation performance in the current study, and the data were shown in Table 4. The baseline median and the baseline absolute deviation were used to represent individual differences in time estimation. No significant difference was found in either the length or the variability of baseline time estimations between the two groups, suggesting that the participants in the two groups do not differ significantly in their time estimation capability. For the organization session, no significant difference was found in either the length or the variability of time estimations of the two experimental groups. For the retrieval session, tagging users tended to exhibit higher variability (average deviation = 3.39) in time estimation than folder users (average deviation = 2.75; $Z = -1.57$, $p = .11$), though no significant difference in the length was found. The result indicated that retrieval tasks tend to demand more cognitive resources for tagging users than for folder users.

Table 4: Time Estimation Performance in Different Sessions

	<i>Tagging^a</i>		<i>Categorizing^a</i>	
	<i>Mdn</i>	<i>Absolute Deviation</i>	<i>Mdn</i>	<i>Absolute Deviation</i>
Baseline	12.3	1.24	11.8	1.04
Organization	17.03	4.68	16.30	4.20
Retrieval	12.73	3.39	12.27	2.75

^a $N = 20$.

4.4. Discussion

The finding that tagging users experienced more mental demand and frustration in organization tasks contradicts the popular belief that “tagging is less cognitively costing” but is consistent with Pak’s study of picture organization tasks (Pak, Pautz, & Iden, 2007). Also using NASA-TLX as the measurement, Pak found that the frustration rating of the tagging group was generally larger than that of the taxonomy group. Comparing our finding with other previous studies, it is interesting to note that users claimed that tagging requires less cognitive effort in interviews (Civan et al., 2009; Quan et al., 2003). When the workload level was measured immediately upon the completion of tasks, as in the current study, however, the opposite was found. The high mental demand may be attributed to the multiple goals that users want to achieve with tags. Recent studies show that tagging users are concerned with the ease of future retrieval and the quality of the organization scheme, like categorization users (Sen et al., 2006; Wash & Rader, 2007). In addition, they have other motivations for tagging, like self-expression (Marlow et al., 2006; Sen et al., 2006). Users create and select tags for multiple goals, which leads to more tags, of different levels of specificity, describing different perspectives of the content. These lead to more mental activity in tagging. The higher frustration of tagging users might be caused by the high overhead required to remember used tags and personal rules for effective and consistent tagging. Past research has found that tagging users worry about the redundancy and inconsistency problems, and sometimes they even give up the multiple tagging function to avoid such problems (Civan et al., 2009; Quan et al., 2003; Wash & Rader, 2007).

For information retrieval tasks, the error rate with the tagging interface was significantly higher than that with the categorization interface. There are several plausible reasons: First, the folder system allows systematic and exhaustive search using a general-to-specific approach, whereas tag systems do not; second, the lack of a clear structure and the difficulty of keeping tagging consistent in the organization session may also contribute to the high error rate in the retrieval session. Although tagging provides more flexible and direct routes for retrieving information, our results imply that it is more prone to errors than hierarchical structured folders. Tagging users also tended to be less stable in the secondary task performance, and they rated the temporal demand of retrieval tasks higher than category users did. It is interesting to compare our finding with those in the Pak et al. (2007) study, which found that tagging users reported a significantly higher level of temporal demand for picture organization tasks, but this effect disappeared in the following retrieval session. A difference between the two studies is that participants in the Pak et al. study had speed requirements for both sessions; in our study, only the retrieval session had speed requirements. It seems that tagging users are more likely to be engaged in a speed/accuracy trade-off than categorization users when they start to have speed constraints, as indicated by the higher error rate and shorter completion time. However, this difference may disappear as users get used to the temporal pressure.

5. EXPERIMENT 2

5.1. Research Questions

In previous research and in our pilot study, tagging users were found to have a strong intention to reuse used tags they created before and to adhere to their own tagging rules if they can remember (Wash & Rader, 2007). Displaying previously used tags as visual aids was found to influence users' tag selection (Binkowski, 2006; Sen et al., 2006). The way in which used tags are displayed is likely to affect how easily users can recognize and find tags they used before and select tags in a more consistent way, yet this influence has not yet been studied. Experiment 2 was designed to examine how users' tagging consistency is influenced by the way previously used tags are displayed. Two hypotheses were proposed.

H1: Visualizing the frequency of tags by font size variations will improve the tagging consistency of individual users.

Font size is found one to be of the most prominent visual attributes perceived by users (Bateman, 2007), and large font sizes are more easily found, recognized, and recalled (Halvey & Keane, 2007; Rivadeneira et al., 2007). Making frequently used tags larger is expected to facilitate the recognition of these tags at a relatively low cost of visual search. Therefore, frequency visualization by font size is expected to increase the tagging consistency without increasing tagging workload.

H2: The visualization of semantic similarities among tags by tag clustering will improves intratagger consistency.

Clustering tags according to their semantic similarity, which can be inferred by co-occurrence similarity or latent semantic analysis (Choy & Lui, 2006; Hassan-Montero & Herrero-Solana, 2006), means that tags close to one another can be inferred to be similar. This may facilitate the cognitive processing of the used tags. Further, users may search for other proper tags in a small area near the first tag they have selected, if they acknowledge the relationship between visual distance and semantic relations. The reduced search scope may decrease the uncertainty in selection and reduce the visual search effort. Therefore, it is predicted that visualizing semantic relationships among tags helps users in stabilizing their tagging patterns without increasing the tagging workload.

5.2. Methodology

Participants. To test the hypotheses, a 2×2 experiment design was used. Forty participants were recruited for Experiment 2 via a similar approach to that used in Experiment 1. The sample consisted of 10 female and 30 male participants. Thirty-one participants were undergraduate students and nine were graduate students. Their ages ranged from 20 to 31 ($M = 22.8$, $SD = 2.03$). Participants had used

computers for an average of 8.7 years ($SD = 3.02$) and used the Internet for an average of 7.2 years ($SD = 2.24$). As required, all of them had experience with tagging. They were randomly assigned to four treatment groups, leaving 10 participants in each group. An examination of age, computer experience, and Internet experience showed no significant differences between the four groups.

Materials. Pictures were used as stimuli in Experiment 2 because they are easy to understand yet contain enough clues for descriptions from multiple aspects. One hundred royalty-free color pictures were picked at random from among pictures tagged as “nature,” “city,” and “people” on Flickr (<http://www.flickr.com>), with the constraint that participants should be able to describe the picture easily without prior knowledge of the scene or event presented. Among them, 20 were stimuli and 80 were filler pictures. Thumbnails of all pictures are shown in the appendix.

Testing systems. The provision of frequency visualization and the provision of semantic-similarity visualization were the two independent variables. The frequency visualization was manipulated in the following way: The font size of tags was weighted according to the number times the tag had been used. The smallest font size of tags was set to 12 pixels. Although there is no available research on the smallest eligible Chinese font size, we found that the size of 12 pixels is acceptable and used widely in Chinese tag clouds on different websites (e.g., myweb.cn.yahoo.com, douban.com, yupoo.com). The largest size was set to 60 pixels, which is generally considered a prominent size. Five levels were set in between, as shown in Table 5.

For a given tag, the font size was determined by the following logarithm function:

$$Current_i = \left\lceil \frac{6 \log(O_i)}{\log(120)} \right\rceil + 1$$

where $Current_i$ is the font size level of the current tag, and O_i is the use frequency of the current tag. The logarithm function was used because the tag distribution is reported to be similar to a power law (Kipp & Campbell, 2006; Mathes, 2004). The logarithm of 120 was used as the denominator because there were 120 tagging

Table 5: Definition of Font Size Levels

Font Size Level (<i>i</i>)	Font Size (px)
1	12
2	20
3	28
4	36
5	44
6	52
7	60

tasks. In the group without frequency visualization, all tags were at the size of 12 pixels.

Design of semantic-similarity visualization depends on how similarity is defined and measured. Co-occurrences among tags were considered to reflect semantic relationships among tags and have been used to infer similarities between tags in various studies (Cattuto et al., 2007; Hassan-Montero & Herrero-Solana, 2006; Michlmayr & Cayzer, 2007; Mika, 2007). We adopted Hassan-Montero and Herrero-Solana's method for semantically clustering of tags. The approach has proven effective in differentiating among main topics and inferring semantic knowledge from the neighbor relationships (Hassan-Montero & Herrero-Solana, 2006).

Tags are represented using the vector-space model. Each tag t_i is considered to be a vector in the document space

$$t_i = (d_{1i}, d_{2i}, d_{3i}, \dots, d_{ni})$$

where d_{ji} is the frequency of the j document being tagged by t_i . The value of d_{ji} is 0 or 1. Each vector is then normalized, denoted as \vec{t}_i .

The similarity between two tags is computed with the cosine measure:

$$\text{cosine}(t_1, t_2) = (t_1 \cdot t_2) / \|t_1\| * \|t_2\|$$

where \cdot denotes dot product among vectors, and $\|t\|$ denotes the length of the vector t . The larger the cosine value is, the more similar the two tags are.

The centroid vector c of a set S of tags is defined as

$$c = \frac{1}{|S|} \sum_{t \in S} t$$

Based on these definitions, general K-means clustering was applied. In the tag cloud, tags from the same cluster are presented in the same line, and clusters of tags are displayed near semantically similar clusters. White (RGB value: FFFFFFFF) and 10% gray background colors are used alternatively for lines to emphasize the clustering effect of tags.

A testing system was developed with JSP and MySQL (5.0). The application was installed on a Lenovo Thinkpad T61P laptop which was equipped a 15.4-in. LCD display set to 1280×800 pixels resolution. The tagging interface is shown in Figure 4. The top of the display shows the picture to be tagged. Below the picture is a text field in which the user can type multiple tags delimited by space. The user can also select tags from a tag cloud shown below the input field consisting of his or her own previously used tags. One of the four different visualization treatments is used for the tag cloud. By clicking a tag in the tag cloud, the tag appears in the text field followed by a space; simultaneously, the background color of the tag in the tag cloud is switched to gray or to white (different from the background color of the line), indicating that the tag has been selected. Typing a used tag in the text

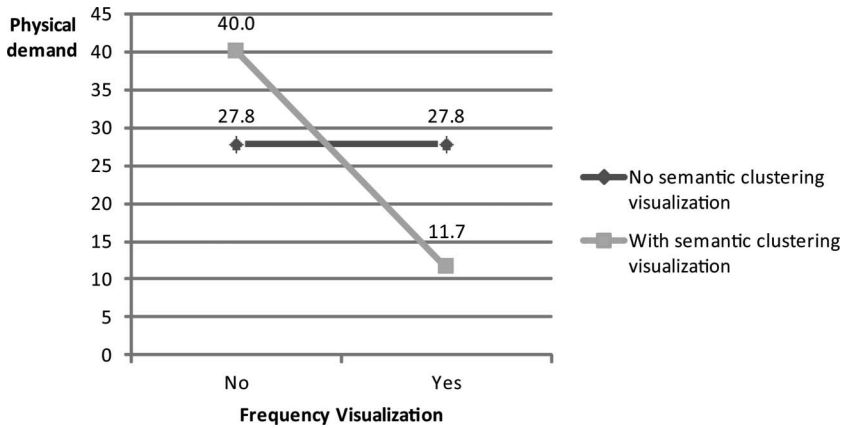


FIGURE 4 Tag visualization treatments in Experiment 2.

field also changes the background color of the corresponding tag in the tag cloud. After clicking the Save button, the user is presented with the next picture to tag.

Dependent variables. The tagging consistency of individual users was measured by the agreement between two indexing sessions: Let A_i and B_i denote the sets of tags that were assigned to the same document i in two sessions. We define the relative overlapping between A_i and B_i by

$$O_i(A, B) = \frac{|A_i \cap B_i|}{|A_i \cup B_i|}.$$

This definition is consistent with the indexing consistency measure proposed by Rodgers (1961), which is widely used in indexing research (Hurwitz, 1969; King & Bryant, 1971; Soergel, 1994). We define the overall overlapping between the two successive runs by the mean of all relative overlappings of this user:

$$O = \frac{\sum_{i=1}^n O_i(A, B)}{n}.$$

In addition, the mental workload perceived by participants was measured with the NASA-TLX scale.

Procedure. The experiment was conducted in the Usability Lab at Tsinghua University. Each participant was tested individually. Before starting the real tasks, a practice session with five trials was conducted to help the participants understand the operation of the system. In the first tagging session, participants were given 60 pictures to tag, including 20 stimuli and 40 filler pictures. The pictures were presented one at a time to the participant at a randomized sequence. The participant was asked to create his own tags from scratch. The tag cloud on the

tagging interface was empty at the beginning and evolved as the participant used more and more tags.

After the first tagging session, the participant was asked to complete the NASA-TLX questionnaire, and then interviewed about his or her tagging patterns and strategies and perception of tagging as a tool for information organization, for about half hour. To further reduce the recency effect, a distracter session was then introduced. The participant was asked to do cumulative addition from 216 (i.e., adding 1, 2, 3 . . . to 216) until the result was larger than 500. Then he was asked to do subtraction from 706 (i.e., subtracting 1, 2, 3, . . . from 706) until the result was smaller than 500. He was asked to report his result loudly after each step, so as to increase the stress level and therefore the disruptive effect. Such mental calculation has been found to be disruptive for human-computer interaction tasks (Gillie & Broadbent, 1989; Kreifeldt & McCarthy, 1981).

Following the interruption, participants had the second tagging session in which they needed to tag 60 pictures, including the 20 stimuli which had been tagged in the first tagging session and another 40 filler pictures. Finally, the participants were interviewed about their perception of the repeated pictures, their tendency to give consistent tags, and their perception of the tag clouds, and 50 Yuan was paid for participation.

5.3. Results

The effects of two visualization techniques and their interaction on users' selection of tags were tested by two-way ANOVA with interaction effect. Data violating the assumptions of ANOVA were tested with the Kruskal-Wallis test. The interaction effects of variables failing to meet the requirements were tested with the nonparametric method proposed by Akritas, Arnold, and Brunner (1997). As shown in Table 6, the examination of the impact of frequency visualization on tagging consistency and workload indicated no significant differences in tagging consistency and the overall workload rating. H1 was not supported. As shown in Table 7, performance, effort, and mental demand were still the three most important factors in the workload, the same as in Experiment 1. The examination of the six subscales of NASA-TLX revealed significant differences in mental demand (Kruskal-Wallis test: $\chi^2 = 4.08$, $p = .04$) and physical demand (Kruskal-Wallis test: $\chi^2 = 4.09$, $p = .04$). Participants tagging with frequency visualization perceived lower levels of physical demand ($M = 22.4$, $SD = 15.84$) than participants tagging without such visualization ($M = 33.9$, $SD = 19.97$), but they also perceived higher level of mental demand ($M = 53.0$, $SD = 15.55$) than the other group ($M = 42.0$, $SD = 16.98$). Furthermore, the interaction effect is significant with physical demand ($\chi^2 = 6.4$, $p = .01$). As shown in Figure 5, participants tagging with frequency visualization ($M = 17.0$, $SD = 11.65$) perceived 57.5% lower physical demand than participants tagging without such visualization ($M = 40.0$, $SD = 16.84$; Kruskal-Wallis test: $\chi^2 = 8.49$, $p = .004$) when the semantically clustered visualization was presented. However, when the semantically clustering visualization was not shown, participants perceived nearly equal levels of physical demand in both groups. In addition, the number of tags participants gave in the two tagging sessions was compared, but no significant difference was found.

Table 6: The Impact of Frequency Visualization on Tagging Consistency and User Workload

	<i>No Frequency Visualization^a</i>		<i>With Frequency Visualization^a</i>		<i>F(1, 36)</i>	<i>p</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
No. of tags in 1st session	47.30	15.35	46.30	19.48	$\chi^2 = 0.12^b$.73
No. of tags in 2nd session	46.40	13.20	46.70	18.39	<.01	.95
Consistency	0.72	0.116	0.69	0.145	0.78	.38
Workload						
Mental demand	42.00	16.98	53.00	15.55	$\chi^2 = 4.09^b$.04
Physical demand	33.90	18.97	22.40	15.84	$\chi^2 = 4.08^b$.04
Temporal demand	32.30	17.86	44.80	20.10	$\chi^2 = 2.93^b$.08
Performance	40.40	19.85	35.87	21.89	0.49	.49
Effort	59.40	19.06	62.00	21.28	$\chi^2 = 0.37^b$.54
Frustration level	22.30	22.45	32.30	24.03	$\chi^2 = 2.14^b$.14
Global	43.00	12.73	46.00	12.35	0.52	.47

^a*N* = 20.

^bKruskal-Wallis test.

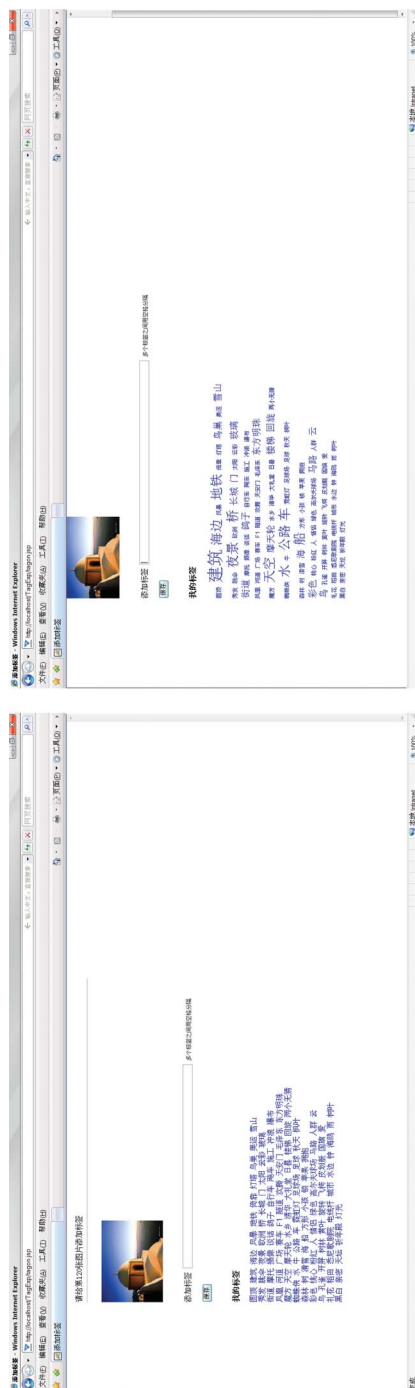
Table 7: Sources-of-Workload Weight of NASA-TLX Subscales in Experiment 2

	<i>Mental Demand</i>	<i>Physical Demand</i>	<i>Temporal Demand</i>	<i>Performance</i>	<i>Effort</i>	<i>Frustration</i>
Sources-of-workload weight	3.15	1.45	1.93	3.50	3.35	1.58
Rank order	3	6	4	1	2	5

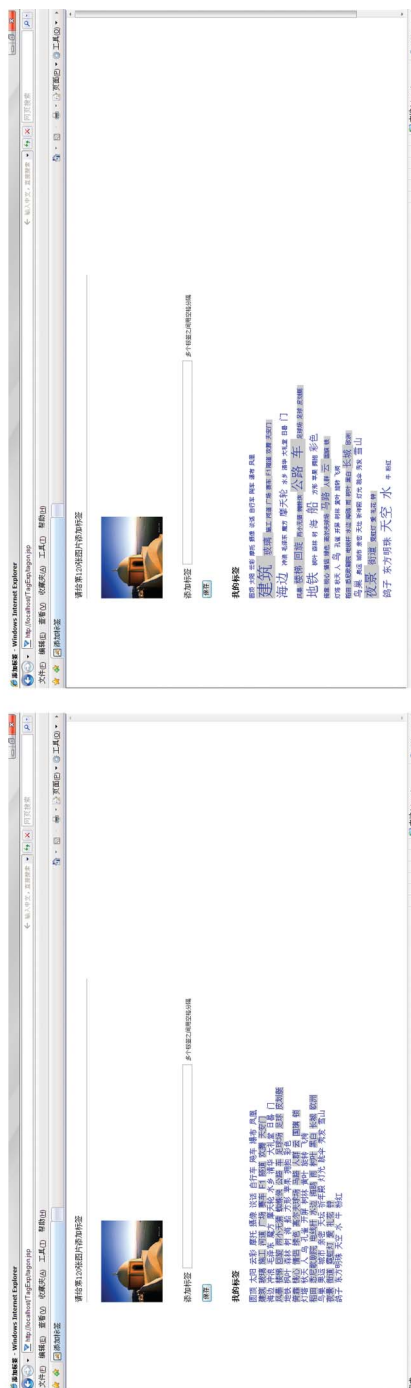
Table 8 shows data for testing the effect of semantic-similarity visualization. A significant difference was found in tagging consistency, $F(1, 36) = 4.0$, $p = .049$. The consistency level of the semantic similarity visualization condition ($M = 0.75$, $SD = 0.13$) is significantly higher than that in the condition without such visualization ($M = 0.67$, $SD = 0.13$). No significant difference was found in the global NASA-TLX score, six subscales, and the number of tags. In summary, presenting semantic similarity with clusters help users improve their tagging consistency without causing a higher workload for tagging. H2 was supported.

5.4. Discussion

The results of Experiment 2 show that semantic-similarity visualization improves tagging consistency significantly without increasing the workload for users. Previous studies found that tagging users have a strong inclination to reuse old tags they created before and to maintain a certain consistency with tagging (Rader & Wash, 2008; Sen et al., 2006), but in practice many fail to do so due to the increased number of tags and rules. With the help of such visual displays, the difficult task of remembering used tags could be partially reduced to a much easier task of searching relevant tags in a small area. By placing related tags close to each other, the links among related tags can be strengthened through visual inferences,



(a) Plain tag cloud (no visualizations)



(b) Only frequency visualization by font size



(c) Both frequency and semantic-similarity visualization

49

FIGURE 5 Interaction between frequency visualization and semantic clustering on physical demand in Experiment 2 (color figure available online).

Table 8: The Impact of Semantic-Similarity Visualization on Tagging Consistency and User Workload

	<i>No Clustering Visualization^a</i>		<i>With Clustering Visualization^a</i>		<i>F(1, 36)</i>	<i>p</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
No. of tags in 1st session	49.4	18.65	44.2	15.92	$\chi^2 = 0.84^b$.36
No. of tags in 2nd session	48.6	17.80	44.6	13.70	0.62	.43
Consistency	0.67	0.126	0.75	0.127	4.0	.049
Workload						
Mental demand	51.1	16.47	43.8	17.17	$\chi^2 = 2.39^b$.12
Physical demand	27.8	18.51	28.5	18.37	$\chi^2 = 0.04^b$.82
Temporal demand	41.9	20.69	35.2	18.78	$\chi^2 = 1.27^b$.26
Performance	32.8	18.28	43.5	22.12	2.81	.10
Effort	63.3	17.32	58.2	22.50	$\chi^2 = 0.24^b$.62
Frustration level	26.8	20.66	27.8	26.58	$\chi^2 = 0.10^b$.74
Global	45.2	10.62	45.9	14.33	0.09	.76

^a*N* = 20.

^bKruskal-Wallis test.

in addition to logical and linguistic inferences. Thus, the explicit visual cue of tag relevancy may facilitate users’ judgments of the relevancy between tags, and users can use their spatial memory to help remember and find used tags. In particular, two participants in the semantic-visualization group mentioned in the interview that they remembered the locations of tags in the cloud and this helped them to refine tags from the cloud.

Frequency visualization was found reducing physical demand significantly. The finding could be explained with Fitts’ Law (Fitts, 1954), which defined the difficulty index of positioning movements as a logarithmic function of the target size in proportion to the distance of movement. The difficulty of hand movements for pointing and selecting big targets was low compared with small targets. This is also in accordance with previous studies on tag cloud features (Halvey & Keane, 2007; Rivadeneira et al., 2007), which found that a large font size allows the frequently used tags to be found more quickly and more easily to be recognized. Of interest, the effect of frequency visualization on physical demand is significant only when the tags are semantically clustered. One possible reason is that users tend to use multiple tags to describe a target from various perspectives, and these tags are often of different popularity. The frequency-visualization improves the search process for popular tags, and other relevant tags could be found in the neighborhood when the tags are semantically clustered. When tags are not semantically clustered, users need to search for other relevant tags in small font sizes, scattered in the cloud, or to type tags directly, and the saving of physical effort of searching for popular tags becomes trivial.

Despite the reduced physical demand, using a large font size for frequently used tags also leads to high level of mental demand. One possible explanation is that the frequency visualization encourages users to search and select tags from the tag cloud, but such tags tend to be broad and inclusive, with high semantic density but poor discrimination capability (Hassan-Montero & Herrero-Solana,

2006). The visual attractiveness of popular tags may obscure useful information that is less popular (Zeldman, 2005). In addition, although there are not necessarily connections between the frequency and similarity, the large font size may create a pop-out effect, grouping tags automatically without focused attention (Stuart et al., 1993). Both of these effects may add to the mental effort required for users to find the proper tags describing the current target specifically.

6. CONCLUSIONS

The results of these experiments are limited in several respects. First, participants in both experiments consisted of mainly young college students, who are well educated, technologically savvy, and experienced with tagging. If the results are to be generalized to people of different ages, education backgrounds, and computer skills, more studies are needed because personal information management style could depend heavily on these variables. Second, the duration and frequency of watching the tag clouds was not measured in Experiment 2. Eye-tracking devices should be used in a future study to measure these variables accurately, thereby allowing the allocation of attention to visual search to be better investigated and its influence on the tag selection to be better studied in this way. Third, although distracter tasks were introduced between the organization session and the retrieval session in Experiment 1 and the interval between the two tagging sessions to reduce the recency effect, the intervals were short. The impact of memory decay over long period on users' performance should be tested with longer intervals (e.g., a week) in future studies.

Despite these limitations, the results of this study are expected to help designers understand the uses of tagging and to shed some light on the design of tagging systems. First, the study provides more understanding of tagging as an information organization and retrieval method from the perspective of individual users, based on both qualitative interviews and laboratory experiments. Although tagging imposes few constraints on users and should reduce the effort to manage information "in theory," in practice this flexibility may increase the difficulty of organizing and retrieving information for untrained users, especially in long-term use. Tagging users tend to experience higher levels of workload for organization tasks, and they make more errors in retrieval tasks. Combined with previous research, which found that users preferred to refind information by location-based browsing, our results seriously challenge the efficiency of tagging as a tool to organize or retrieve information. However, the improved memory of the content suggests that tagging encourages encoding information from multiple perspectives and deeper semantic processing of the content. In conditions where the information is self-generated and the amount of information is limited, such as in personal or enterprise information systems, treating tagging as an annotation tool may be more appropriate than as an alternative for categorization or classification. Thus, the question to be answered is no longer "Will tagging substitute for categorization?" but "How can tagging provide extra information in addition to categorization?" Related research questions may include the following: "How should designers communicate the different functions of categorization and

tagging to users when the system embodies both functions?" and "How should the system interface be designed to encourage users to give more particular tags describing the content rather than general and broad category tags?" Furthermore, the finding that tagging consistency can be influenced by the presentation of suggested tags provides implications for designers. Maintaining a certain consistency in tagging was found to be a common intention of tagging users, but they often fail to do so due to the high flexibility of tagging and the lack of supporting tools. It is recommended that semantic-similarity visualization is employed to present frequently used tags so that users can utilize visual cues to refind used tags relevant to the current task. Frequency visualization is recommended to be used in combination with semantic-similarity visualization to reduce physical load on users.

The scope of the study is limited to the scenario in which individual users manage a particular type of information, such as personal photos, documents, and other files and the experiment setting simulate a single-computer scenario. It would be interesting for future research to investigate the scenario in which users need to organize and retrieve information of mixed types (e.g., search for relevant photos, e-mails, and web pages to prepare a project report) and the scenario in which users store and organize information on multiple computers. Both scenarios have practical implications and present new challenges for effectiveness and efficiency of information organization schemes (Beale & Edmondson, 2007; Ravasio et al., 2004). The scenario of information management can be further extended to multiuser cases where collaboration and sharing is emphasized. In addition to cognitive issues, social factors need to be considered in this scenario. Especially in communities with delimited members where users play different social roles, such as enterprise information systems and collaborative learning systems, people are cognizant that their tags have social values, and this awareness consequently influences their selection of tags (Thom-Santelli & Muller, 2007). It would be interesting for future research to investigate how this change will influence the strategy and performance for information organization and retrieval tasks.

REFERENCES

- Abrams, D., Baecker, R., & Chignell, M. (1998). Information archiving with bookmarks: Personal Web space construction and organization. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 41–48). Los Angeles, CA: ACM Press.
- Adar, E., Karger, D., & Stein, L. A. (1999). Haystack: Per-user information environments. *Proceedings of the Eighth International Conference on Information and Knowledge Management* (p. 413–422). New York, US: ACM New York.
- Agarawala, A., & Balakrishnan, R. (2006). Keepin' it real: Pushing the desktop metaphor with physics, piles and the pen. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1283–1292.
- Ahn, J., Brusilovsky, P., & Farzan, R. (2005). Investigating users' needs and behavior for social search. *Proceedings of Workshop on New Technologies for Personalized Information Access (PIA 2005)* (pp. 1–12). Edinburgh, Scotland, UK. Retrieved from http://www.witi.cs.unimagdeburg.de/iti_dke/Events/2005/pia2005/docs/AhnBruFar05.pdf (last accessed July 2011)

- Akritis, M., Arnold, S., & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92, 258–265.
- Ames, M., & Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. *Proceedings of the 2007 SIGCHI Conference on Human Factors in Computing Systems*, 971–980.
- Anderson, J. (2000). *Learning and memory: An integrated approach*. New York, NY: Wiley.
- Anderson, J., & Reder, L. (1979). An elaborative processing explanation of depth of processing. In L. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 385–403). Hillsdale, NJ: Erlbaum.
- Barreau, D., & Nardi, B. A. (1995). Finding and reminding: File organization from the desktop. *ACM SIGCHI Bulletin*, 27(3), 39–43.
- Barsalou, L. (1991). Deriving categories to achieve goals. *The Psychology of Learning and Motivation: Advances in Research and Theory*, 27, 1–64.
- Bateman, S. (2007). Collaborative Tagging: Folksonomy, Metadata, Visualization, E-Learning (Master thesis). University of Saskatchewan. Retrieved from http://library2.usask.ca/theses/available/etd-12112007-221606/unrestricted/scott_bateman_thesis.pdf (last accessed July 2011)
- Beale, R., & Edmondson, W. (2007). Multiple carets, multiple screens and multi-tasking: new behaviours with multiple computers. In *Proceedings of the 21st British CHI Group Annual Conference on HCI 2007: People and Computers XXI* (Vol. 1, pp. 55–64). Swinton, UK: British Computer Society.
- Binkowski, P. (2006). *The effect of social proof on tag selection in social bookmarking applications*. Unpublished master's thesis, University of North Carolina at Chapel Hill.
- Boardman, R., & Sasse, M. A. (2004). "Stuff goes into the computer and doesn't come out": A cross-tool study of personal information management. *Proceedings of the 2004 SIGCHI conference on Human Factors in Computing Systems*. (pp. 583–590). New York, US: ACM New York.
- Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53, 225–250.
- Bower, G. H., Clark, M. C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorical word lists. *Journal of Verbal Learning and Verbal Behaviors*, 8, 323–343.
- Bradshaw, G., & Anderson, J. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior*, 21, 165–174.
- Braly, M., & Froh, G. (2006). *Social bookmarking in the Enterprise*. Poster session presented at the 17th Annual ASIS&T SIG/CR Classification Research Workshop. Austin, TX, March.
- Bransford, J., Franks, J., Morris, C., & Stein, B. (1979). Some general constraints on learning and memory research. In L. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 331–354). Hillsdale, NJ: Erlbaum.
- Brooks, C. H., & Montanez, N. (2006). Improved annotation of the blogosphere via autotagging and hierarchical clustering. *Proceedings of the 15th International Conference on World Wide Web*. New York, US: ACM New York. Retrieved from <http://portal.acm.org/citation.cfm?doid=1135777.1135869> (last accessed July 2011)
- Budiu, R., Piroli, P., & Hong, L. (2009). *Remembrance of things tagged: How tagging effort affects tag production and human memory*. Paper presented at the Proceedings of the 27th International Conference on Human Factors in Computing Systems, Boston, MA. pp. 615–624.
- Campbell, K. E. (2006). A phenomenological framework for the relationship between the semantic Web and user-centered tagging systems. In J. Furnas & J. T. Tennis (Eds.), *Proceedings 17th Workshop of the American Society for*

- Information Science and Technology Special Interest Group in Classification Research, November 4, 2006. Austin, Texas. Retrieved from <http://arizona.openrepository.com/arizona/bitstream/10150/105357/1/campbell.pdf>
- Casali, J. G., & Wierwille, W. W. (1984). On the measurement of pilot perceptual workload: A comparison of assessment techniques addressing sensitivity and intrusion issues. *Ergonomics*, 27, 1033–1050.
- Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V. D. P., Loreto, V., Hotho, A., et al. (2007). Network properties of folksonomies. *AI Communications Journal*, 20, 245–262.
- Choy, S., & Lui, A. (2006). Web information retrieval in collaborative tagging systems. In *Proceedings of web intelligence* (pp. 352–355). Washington, DC: IEEE Computer Society.
- Civan, A., Jones, W., Klasnja, P., & Bruce, H. (2009). Better to organize personal information by folders or by tags?: The devil is in the details. *Proceedings of the American Society for Information Science and Technology*, 45(1), 1–13.
- Cook, J. R. (1991). *The cognitive and social factors in the design of computerized tasks*. Unpublished doctoral dissertation, Purdue University, West Lafayette, IN.
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671–684.
- Dourish, P., Edwards, W. K., LaMarca, A., Lamping, J., Petersen, K., Salisbury, M., et al. (2000). Extending document management systems with user-specific active properties. *ACM Transactions on Information Systems (TOIS)*, 18, 140–170.
- Dumais, S. T., & Jones, W. P. (1985). A comparison of symbolic and spatial filing. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 127–130). New York, US: ACM New York.
- Fertig, S., Freeman, E., & Gelernter, D. (1996). Lifestreams: An alternative to the desktop metaphor. In *Conference companion on human factors in computing systems: Common ground* (pp. 410–411). New York, US: ACM New York.
- Fitts, P. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47, 381–391.
- Freeman, L., Romney, A., & Freeman, S. (1987). Cognitive structure and informant accuracy. *American Anthropologist*, 89, 310–325.
- Gillie, T., & Broadbent, D. (1989). What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological Research*, 50, 243–250.
- Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32, 198–208.
- Gonçalves, D., & Jorge, J. A. (2004). Describing documents: What can users tell us? In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (pp. 247–249). New York, US: ACM New York.
- Grudin, J. (2006). Enterprise knowledge management and emerging technologies. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences* (Vol. 3, p. 57). Washington, DC: IEEE Computer Society.
- Guy, M., & Tonkin, E. (2006). Folksonomies: Tidying up tags. *D-Lib Magazine*, 12(1). Retrieved from <http://www.dlib.org/dlib/january06/guy/01guy.html> (last accessed July 2011)
- Halvey, M., & Keane, M. (2007). An assessment of tag presentation techniques. *Proceedings of the 16th International Conference on World Wide Web* (pp. 1313–1314). New York, US: ACM New York. Retrieved from <http://portal.acm.org/citation.cfm?doid=1242572.1242826>; (last accessed July 2011)
- Hart, S. G., Mcpherson, D., & Loomis, L. L. (1978). Time estimation as a secondary task to measure workload: Summary of Research. *Proceedings of the 14th Annual Conference on Manual Control* (pp. 693–712). Washington, D. C.: NASA.

- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*, 1, 139–183.
- Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, 108, 356–388.
- Hassan-Montero, Y., & Herrero-Solana, V. (2006). *Improving tag-clouds as visual information retrieval interfaces*. Paper presented at the International Conference on Multidisciplinary Information Sciences & Technologies, Mérida, Spain, October 25–28.
- Hurwitz, F. I. (1969). A study of indexer consistency. *American Documentation*, 20(1), 92–94.
- Jacko, J., & Salvendy, G. (1996). Hierarchical menu design: Breadth, depth, and task complexity. *Perceptual and Motor Skills*, 82, 1187–1201.
- Janecek, P. (2007). Faceted classification in web information architecture A framework for using semantic web tools The Authors Mohammad Nasir Uddin, Department of Library and Information Science, University of Rajshahi, Rajshahi, Bangladesh. *The Electronic Library*, 25, 219–233.
- Jones, W. P., & Dumais, S. T. (1986). The spatial metaphor for user interfaces: experimental tests of reference by location versus name. *ACM Transactions on Information Systems (TOIS)*, 4(1), 42–63.
- Jones, W., Dumais, S., & Bruce, H. (2002). Once found, what then? A study of “keeping” behaviors in the personal use of Web information. *Proceedings of the American Society for Information Science and Technology*, 39(1), 391–402.
- Jones, W., Phuwanartnarak, A. J., Gill, R., & Bruce, H. (2005). Don’t take my folders away!: Organizing personal information to get things done. In *CHI ’05 Extended Abstracts on Human Factors in Computing Systems*. Retrieved from <http://portal.acm.org/citation.cfm?id=1056808.1056952> (last accessed July 2011)
- Kiger, J. I. (1984). The depth/breadth tradeoff in the design of menu-driven interfaces. *International Journal of Man-Machine Studies*, 20, 201–203.
- King, D., & Bryant, E. (1971). *The evaluation of information services and products*. Washington, DC: Information Resources Press.
- Kipp, M., & Campbell, D. (2006). Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. *Proceedings of the American Society for Information Science and Technology*, 43(1), 1–18.
- Kreifeldt, J., & McCarthy, M. (1981). Interruption as a test of the user-computer interface. *Proceedings of the 17th Annual Conference on Manual Control*. California Institute of Technology (pp. 655–667). Jet Propulsion Laboratory, California Institute of Technology, JPL Publication 81–95.
- Kwasnik, B. H. (1991). The importance of factors that are not document attributes in the organization of personal documents. *Journal of Documentation*, 47, 389–398.
- Lansdale, M. (1988). The psychology of personal information management. *Applied Ergonomics*, 19(1), 55–66.
- Lansdale, M. W. (1991). Remembering about documents: Memory for appearance, format, and location. *Ergonomics*, 34, 1161–1178.
- Leonard, L. (1975). *Inter-indexer consistency and retrieval effectiveness: measurement of relationships*. Unpublished master’s thesis, University of Illinois at Urbana-Champaign.
- Li, R., Bao, S., Yu, Y., Fei, B., & Su, Z. (2007). Towards effective browsing of large scale social annotations. *Proceedings of the 16th international conference on World Wide Web* (pp. 943–952). New York, US: ACM New York. Retrieved from <http://portal.acm.org/citation.cfm?id=1242700> (last accessed July 2011)
- Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science*, 48, 40–54.

- Liu, Y., & Wickens, C. (1988). *Patterns of task interference when human functions as a controller or a monitor*. Paper presented at the Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics, Beijing and Shenyang, China.
- Liu, Y., & Wickens, C. D. (1994). Mental workload and cognitive task automaticity: An evaluation of subjective and time estimation metrics. *Ergonomics*, 37, 1843–1854.
- Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55, 291–300.
- Malone, T. W. (1983). How do people organise their desks? Implications for the design of office information systems. *ACM Transactions on Office Information Systems*, 1(1), 99–112.
- Mander, R., Salomon, G., & Wong, Y. Y. (1992). A “pile” metaphor for supporting casual organization of information. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 627–634.
- Mandler, G. (1967). Organization and memory. *The Psychology of Learning and Motivation*, 1, 327–372.
- Mandler, J. M., Seegmiller, D., & Day, J. (1977). On the coding of spatial information. *Memory & Cognition*, 5(1), 10–16.
- Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*. pp. 31–40.
- Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Computer Mediated Communication (LIS590CMC). Retrieved from <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html> (last accessed July 2011)
- Michlmayr, E., & Cayzer, S. (2007). *Learning user profiles from tagging data and leveraging them for personal (ized) information access*. Paper presented at the the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference, Banff, Alberta, Canada, May 2007.
- Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1), 5–15.
- Miller, D. P. (1981). The depth/breadth tradeoff in hierarchical computer menus. *Proceedings of the Human Factors Society*, 296–300.
- Nardi, B., Anderson, K., & Erickson, T. (1995). Filing and finding computer files. *Proceedings of the East-West HCI*. pp. 162–179.
- Nardi, B., & Barreau, D. (1997). “Finding and reminding” revisited: Appropriate metaphors for file organization at the desktop. *SIGCHI Bulletin*, 29(1). Retrieved from <http://homepages.cwi.nl/~steven/sigchi/bulletin/1997.1/nardi.html> (last accessed July 2011)
- Naumann, F., & Rolker, C. (2000). Assessment methods for information quality criteria. *Proceedings of 5th International Conference on Information Quality*, 148–162.
- Nielsen, M. (2007). Functionality in a second generation tag cloud (Master thesis). Gjøvik University College. Retrieved from <http://www.hig.no/content/download/9053/122120/file/Nielsen%20-%20Functionality%20in%20a%20second%20generation%20tag%20cloud.pdf> (last accessed July 2011)
- Pak, R., Pautz, S., & Iden, R. (2007). Information organization and retrieval: A comparison of taxonomical and tagging systems. *Cognitive Technology*, 12(1), 31–44.
- Pak, R., Rogers, W. A., & Fisk, A. D. (2006). Spatial ability subfactors and their influences on a computer-based information search task. *Human Factors*, 48(1), 154–165.
- Pak, R., Pautz, S., & Iden, R. (2007). Information organization and retrieval: A comparison of taxonomical and tagging systems. *Cognitive Technology*, 12, 31–44.

- Quan, D., Bakshi, K., Huynh, D., & Karger, D. (2003). User interfaces for supporting multiple categorization. In M. Rauterberg, M. Menozzi, & J. Wesson (Eds.), *Human-computer interaction: INTERACT '03* (pp. 228–235). Amsterdam, the Netherlands: IOS Press.
- Rader, E., & Wash, R. (2008). Influences on tag choices in del.icio.us. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, 239–248.
- Ravasio, P., Schär, S. G., & Krueger, H. (2004). In pursuit of desktop evolution: User problems and practices with modern desktop systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 11, 156–180.
- Rivadeneira, A., Gruen, D., Muller, M., & Millen, D. (2007). Getting our head in the clouds: toward evaluation studies of tagclouds. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 995–998). New York, US: ACM New York. Retrieved from <http://portal.acm.org/citation.cfm?doid=1240624.1240775> (last accessed July 2011)
- Robertson, G., Czerwinski, M., Larson, K., Robbins, D. C., Thiel, D., & Van Dantzich, M. (1998). Data mountain: Using spatial memory for document management. *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, 153–162.
- Rodgers, D. (1961). *A study of inter-indexer consistency*. Washington, DC: General Electric Co.
- Schrammel, J., Leitner, M., & Tscheligi, M. (2009). Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. *Proceedings of the 27th international conference on Human factors in computing systems*. New York, US: ACM New York. Retrieved from <http://portal.acm.org/citation.cfm?id=1519010> (last accessed July 2011)
- Seagull, F. J., & Walker, N. (1992). The effects of hierarchical structure and visualization ability on computerized information retrieval. *International Journal of Human-Computer Interaction*, 4, 369–385.
- Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., et al. (2006). Tagging, communities, vocabulary, evolution. *Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work*. pp. 181–190.
- Shirky, C. (2005). *Ontology is Overrated: Categories, Links, and Tags*. Economics & Culture, Media & Community. Retrieved from http://shirky.com/writings/ontology_overrated.html (last accessed July 2011)
- Sinha, R. (2005). *A cognitive analysis of tagging*. http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html (last accessed March 2005)
- Snowberry, K., Pakinson, S. R., & Sisson, N. (1983). Computer display menus. *Ergonomics*, 26, 699–712.
- Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science*, 45, 589–599.
- Stanney, K., & Salvendy, G. (1995). Information visualization: Assisting low spatial individuals with information access tasks through the use of visual mediators. *Ergonomics*, 38, 1184–1198.
- Stuart, G. W., Bossomaier, T. R. J., & Johnson, S. (1993). Preattentive processing of object size: Implications for theories of size perception. *Perception*, 22, 1175–1193.
- Thom-Santelli, J., & Muller, M. (2007). *The wisdom of my crowd: Motivation and audience in enterprise social tagging*. Poster presented at GROUP, Sanibel Island, FL.
- Toda, H., Kataoka, R., & Oku, M. (2007). Search result clustering using informatively named entities. *International Journal of Human-Computer Interaction*, 23, 3–23.
- Tulving, E. (1962). Subjective organization in free recall of unrelated words. *Psychological Review*, 69, 344–354.

- Veres, C. (2006). The language of folksonomies: What tags reveal about user classification. *Lecture Notes in Computer Science*, 3999, 58–69.
- Vidulich, M., & Tsang, P. (1985). *Assessing subjective workload assessment: A comparison of SWAT and the NASA-Bipolar methods*. Paper presented at the Human Factors Society 20th Annual Meeting, Santa Monica, CA, September.
- Wash, R., & Rader, E. (2007). Public bookmarks and private benefits: An analysis of incentives in social computing. *Proceedings of the 2007 Annual Meeting of American Society for Information Science and Technology*, 1–13.
- Whittaker, S., & Sidner, C. (1996). Email overload: Exploring personal information management of email. *Proceedings of the 1996 SIGCHI Conference on Human Factors in Computing Systems: Common Ground*, 276–283.
- Wu, X., Zhang, L., & Yu, Y. (2006). Exploring social annotations for the semantic web. *Proceedings of the 15th International Conference on World Wide Web*. pp. 417–426.
- Xu, Z., Fu, Y., Mao, J., & Su, D. (2006). Towards the semantic web: Collaborative tag suggestions. *Proceedings of Workshop on Collaborative Web Tagging*, Edinburgh, Scotland: Retrieved from <http://www.semanticmetadata.net/hosted/taggingws-www2006-files/13.pdf> (last accessed July 2011)
- Zaphiris, P. (2000). Depth vs breadth in the arrangement of web links. *Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society*, 139–144.
- Zeldman, J. (2005, June, 5). *Remove forebrain and serve: Tag clouds II*. <http://www.zeldman.com/daily/0505a.shtml> (last accessed July 2011)

APPENDIX

Pictures Used in Experiment 2

Pictures used in Session 1:







Pictures used in Session 2:

